ACADEMIC PRESS

J. Math. Anal. Appl. 273 (2002) 93–111

Journal of MATHEMATICAL ANALYSIS AND APPLICATIONS

www.academicpress.com

Nonatomic total rewards Markov decision processes with multiple criteria

Eugene A. Feinberg [a,*,1] and Aleksey B. Piunovskiy [b]

[a] *Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600, USA*

[b] *Department of Mathematical Sciences, The University of Liverpool, M & O Building, Liverpool, L69 7ZL, UK*

Received 24 January 2001

Submitted by S.M. Meerkov

Abstract

We consider a Markov decision process with an uncountable state space for which the vector performance functional has the form of expected total rewards. Under the single condition that initial distribution and transition probabilities are nonatomic, we prove that the performance space coincides with that generated by nonrandomized Markov policies. We also provide conditions for the existence of optimal policies when the goal is to maximize one component of the performance vector subject to inequality constraints on other components. We illustrate our results with examples of production and financial problems.

1. Introduction

This paper deals with a Markov decision process (MDP) with Borel state and action spaces and multiple performance criteria. Each criterion has the form of expected total rewards. A standard natural approach to such problems is to

* Corresponding author.
E-mail addresses: eugene.feinberg@sunysb.edu (E.A. Feinberg), piunov@liverpool.ac.uk (A.B. Piunovskiy).

[1] Research of this coauthor was partially supported by NSF Grant DMI-9908258.

0022-247X/02/$ – see front matter
PII: S0022-247X(02)00213-5

optimize one of these criteria under inequality constraints on other criteria. The well-known phenomenon for problems with constraints is that optimal strategies, if they exist, may be randomized and nonrandomized optimal strategies may not exist; see Altman [1], Frid [12], or Piunovskiy [18].

An MDP is called nonatomic if all transition probabilities are nonatomic and a nonatomic initial distribution is fixed. Feinberg and Piunovskiy [10] proved that if a nonatomic MDP satisfies continuity and compactness conditions then there exists an optimal nonrandomized Markov policy for a multiple criterion problem with constraints. This result was established there as a corollary of the following fact: the performance set for nonrandomized Markov policies coincides with the performance set for all policies in an MDP satisfying continuity and compactness conditions. Continuity and compactness conditions were essential for the proofs in Feinberg and Piunovskiy [10].

In this paper we prove that the performance set for nonrandomized Markov policies coincides with the performance set for all policies for an arbitrary nonatomic MDP with a vector criterion of expected total rewards. Our proofs use Liapunov's theorem. We recall that Borel space $(\Omega, \mathcal{B}(\Omega))$ is a measurable space isomorphic to a Polish space; see Bertsekas and Shreve [5], Dynkin and Yushkevich [8], or Feinberg and Piunovskiy [10] for details.

Liapunov's theorem (Barra [4, p. 218]). *Let* $\{P_1, P_2, \ldots, P_n\}$ *be a finite collection of nonatomic probability distributions on the Borel space* $(\Omega, \mathcal{B}(\Omega))$. *Then, for each random variable* $\hat{\pi}(x)$ *with values in* $[0, 1]$, *there exists a measurable set* $\Gamma \in \mathcal{B}(X)$ *such that*

$$P_n(\Gamma) = \int_X P_n(dx)\,\hat{\pi}(x), \quad n = 1, 2, \ldots, N.$$

We describe the model and formulate the main result, Theorem 2.1, in Section 2. We prove Theorem 2.1 in Section 4. We provide a preliminary proof for a particular case of a one-step problem in Section 3.

Theorem 2.1, that states that for any policy there exists a nonrandomized Markov policy with the same performance vector, assumes only the condition that the initial and transition probabilities do not have atoms. It does not require any compactness and continuity conditions and therefore does not guarantee the existence of optimal policies. Section 5 provides sufficient conditions for the existence of optimal policies for problems with multiple criteria and constraints. We consider two types of conditions: Condition 5.2(a) (i) and (ii). In both cases, one-step reward functions can be unbounded. Unboundedness of rewards is important because holding costs in inventory, production, and queuing problems may tend to ∞ with the increasing inventory/queue size. Our Condition 5.2(a) (ii) is similar to the condition considered for discounted problems in the recent paper by Hernandez-Lerma and Gonzalez-Hernandez [13] with the following major

difference: it was assumed in [13] that the one-step reward function associated with the objective function is sup-compact. We assume in Condition 5.2(a) (ii) that there exist sup-compact linear combinations of one-step reward functions. In particular, any one-step reward function, associated either with the objective function or a constraint, can be sup-compact.

In Section 6 we consider two particular examples: a production/inventory application and a financial application. In particular, our Example 6.1 demonstrates the natural situation when none of one-step reward functions is sup-compact but there is a linear combination of these functions which is sup-compact. We prove the existence of optimal nonrandomized Markov policies in Examples 6.1 and 6.2 under much broader conditions than in similar examples in [10].

2. Model description and main result

We consider a Markov decision process (MDP) $\{X, A, A(\cdot), p, r\}$, where

(i) X is a Borel state space;
(ii) A is a Borel action space;
(iii) $A_t(x)$ are sets of actions available at states $x \in X$ at epochs $t = 0, 1, \ldots$; it is assumed that for each t the graph $\mathrm{Gr}(A_t) = \{(x, a)\colon\ x \in X,\ a \in A(x)\}$ is a measurable subset of $A \times X$ and there exists a measurable mapping $\varphi : X \to A$ with $\varphi(x) \in A_t(x)$ for all $x \in X$. (In other words, the graphs of A_t are measurable and multifunctions $x \to A_t(x)$ can be uniformized; see [8, 20]);
(iv) $p_t(dy|x, a)$ are measurable transition probabilities from $X \times A$ to X at steps $t = 1, 2, \ldots$;
(v) $r_t(x, a) = (r_t^1, r_t^2, \ldots, r_t^N)$ are N-dimensional vectors of measurable rewards with values in $[-\infty, \infty]$ at steps $t = 0, 1, \ldots$, where N is a positive integer and $(x, a) \in X \times A$.

As usually, a policy π is a sequence of measurable transition probabilities $\pi_t(da|h_t)$ concentrated on the sets $A_t(x_t)$, where $h_t = x_0, a_0, \ldots, a_{t-1}, x_t$ is the observed history. Δ is the set of all policies. If transition probabilities π_t depend only on the current time and the current state, i.e., $\pi_t(\cdot|h_t) = \pi_t(\cdot|x_t)$ for all $t = 0, 1, \ldots$, then the policy π is called randomized Markov. If the measure π_t, for all $t = 0, 1, \ldots$, is concentrated at the point $\varphi_t(x_t) \in A_t(x_t)$, then the policy is called nonrandomized Markov and is denoted by φ. Δ^M is the set of all nonrandomized Markov policies.

According to Ionescu–Tulcea theorem [8] an initial distribution μ on X and a policy π define a unique probability measure P_μ^π on the space of trajectories $H_\infty = (X \times A)^\infty$ which is called strategic measure. We denote by E_μ^π expectations with respect to P_μ^π. Since the initial distribution μ is fixed, the index

μ is omitted usually. $\mathcal{D}$ is the set of all strategic measures with the initial measure μ, $\mathcal{D}^M$ is the set of all strategic measures generated by nonrandomized Markov policies and the initial measure μ.

For a Borel space $(\Omega, \mathcal{B}(\Omega))$, we denote by $\mathcal{P}(\Omega)$ the set of all probability measures on it. We also denote by $\mathcal{M}(\Omega)$ the minimal σ-field on $\mathcal{P}(\Omega)$ with respect to which all functions $P(E)$ are measurable for every fixed $E \in \mathcal{B}(\Omega)$. Then $(\mathcal{P}(\Omega), \mathcal{M}(\Omega))$ is a Borel space; see Dynkin and Yushkevich [8, Appendix 5]. We also notice that $\mathcal{P}(\Omega)$ is a convex subset of the linear space of all signed finite measures on $(\Omega, \mathcal{B}(\Omega))$.

In what follows, $C^+ = \max\{C, 0\}$, $C^- = \min\{C, 0\}$;

$$R^n_+(P^\pi) = E^\pi\left[\sum_{t=0}^{\infty}\bigl(r^n_t(x_t, a_t)\bigr)^+\right],$$

$$R^n_-(P^\pi) = E^\pi\left[\sum_{t=0}^{\infty}\bigl(r^n_t(x_t, a_t)\bigr)^-\right],$$

$$R^n(P^n) = R^n_+(P^\pi) + R^n_-(P^\pi),$$

where throughout this paper $+\infty - \infty = -\infty$. The performance of a policy π is evaluated by a vector

$$\mathbf{R}(P^\pi) = \bigl(R^1(P^\pi), R^2(P^\pi), \ldots, R^N(P^\pi)\bigr). \tag{1}$$

Let us introduce performance spaces

$$\mathcal{V} = \bigl\{\mathbf{R}(P^\pi),\ \pi \in \Delta\bigr\}, \qquad \mathcal{V}^M = \bigl\{\mathbf{R}(P^\varphi),\ \varphi \in \Delta^M\bigr\}.$$

Infinite values $R^n(P^\pi) = \pm\infty$ can be obtained for some policies. Let $\mathbb{R}^N$ be the Euclidean space of N-dimensional vectors with finite coordinates. Then subsets $\mathcal{V} \cap \mathbb{R}^N$ and $\mathcal{V}^M \cap \mathbb{R}^N$ consist of vectors with finite elements.

We assume that the following condition is satisfied throughout this paper.

Condition 2.1. *The initial measure $\mu(\cdot)$ is nonatomic and for every triple (t, x, a), where $t = 0, 1, 2, \ldots$, $x \in X$, and $a \in A_t(x)$, the transition measure $p_{t+1}(\cdot\,|x, a)$ is nonatomic.*

Theorem 2.1. *Under Condition* 2.1, $\mathcal{V} = \mathcal{V}^M$.

We prove Theorem 2.1 in Section 4. To do it, we prove it first for a one-step problem in Section 3. Since $\mathcal{V} \supseteq \mathcal{V}^M$, the equality in Theorem 2.1 is equivalent to $\mathcal{V} \subseteq \mathcal{V}^M$. We can double the dimension of vectors $\mathbf{R}$ by considering separately positive and negative parts of reward functions. If Theorem 2.1 holds for the new model, it holds for the original model. Thus, it is sufficient to prove Theorem 2.1 only for nonnegative reward functions. Thus, we assume without loss of generality everywhere until the end of Section 4 that $r_t(\cdot) \geqslant 0$ for all t.

The set of strategic measures $\mathcal{D}$ is a measurable convex subset of $\mathcal{P}(H_\infty)$; see Dynkin and Yushkevich [8]. Since functions r_t^n are nonnegative, $\mathbf{R}$ is an affine mapping. Therefore, $\mathcal{V} = \mathbf{R}(\mathcal{D})$ is convex.

3. One-step model

Suppose that $r_t^n(\cdot) = 0$ by $t \geqslant 1$ for all $n = 1, 2, \ldots, N$. To put it differently, the control process ends after we chose the stochastic kernel π_0. The index $t = 0$ is omitted everywhere in this section. The set of all nonrandomized Markov policies for this model coincides with the set of all nonrandomized policies.

Lemma 3.1. *In the one-step model with the nonatomic measure μ the set $\mathcal{V}^M \cap \mathbb{R}^N$ is convex.*

Proof. Let us fix two arbitrary nonrandomized policies $\varphi^1(x)$ and $\varphi^2(x)$ such that the vectors $\mathbf{R}(P^{\varphi^1}) \neq \mathbf{R}(P^{\varphi^2})$ are finite. For an arbitrary $\alpha \in]0, 1[$, we consider

$$v = \alpha \mathbf{R}\big(P^{\varphi^1}\big) + (1-\alpha)\mathbf{R}\big(P^{\varphi^2}\big).$$

To prove Lemma 3.1, it is sufficient to show that $v = \mathbf{R}(P^{\varphi})$ for some (nonrandomized) policy φ.

We define an MDP which elements, except the sets of available actions, coincide with the elements of the original MDP. Let the sets of available actions in the new MDP be $\tilde{A}(x) \triangleq \{\varphi^1(x), \varphi^2(x)\}$, $\tilde{A}(x) \subseteq A(x)$, $x \in X$. Let $\tilde{\Delta}$ be the set of all policies in this model. Then the performance set $\widetilde{\mathcal{V}} = \{\mathbf{R}(P^{\pi}),\ \pi \in \tilde{\Delta}\}$ is convex. Therefore, $v = \mathbf{R}(P^{\pi})$ for some policy π in the new MDP. Let $\hat{\pi}(x) \triangleq \pi(\varphi^1(x)|x)$.

Any nonrandomized policy φ in the new MDP has the form

$$\varphi(x) = \begin{cases} \varphi^1(x), & \text{if } x \in \Gamma, \\ \varphi^2(x), & \text{if } x \in X \setminus \Gamma, \end{cases}$$

for some $\Gamma \in \mathcal{B}(X)$. Our goal is to construct Γ such that $\mathbf{R}(P^{\varphi}) = \mathbf{R}(P^{\pi})$.

We define a finite partition $\{Y_i\}$ of X by

$$\begin{aligned} Y_i \triangleq \Big\{ x \in X\colon\ & r^1\big(x, \varphi^1(x)\big) \sim_1 r^1\big(x, \varphi^2(x)\big), \\ & r^2\big(x, \varphi^1(x)\big) \sim_2 r^2\big(x, \varphi^2(x)\big), \ldots, \\ & r^N\big(x, \varphi^1(x)\big) \sim_N r^N\big(x, \varphi^2(x)\big) \Big\}, \end{aligned} \tag{2}$$

where $\sim_n \in \{>, <, =\}$, $n = 1, \ldots, N$. Each set Y_i is measurable. We shall construct Γ in the form $\Gamma = \bigcup_i \Gamma_i$, where each Γ_i is a measurable subset of Y_i. To do it, we need to construct all Γ_i.

First, if $\mu(Y_i)=0$, we define Γ_i as an arbitrary measurable subset of Y_i. So, we consider only Y_i for which $\mu(Y_i)>0$. Second, we start with a set Y_i such that all inequalities in (2) are strict inequalities. In other words, $\sim_n \in \{>,<\}$, $n=1,\dots,N$, in the definition of a set Y_i.

Let us introduce the collection of nonatomic probability measures on the Borel space $(Y_i,\mathcal{B}(Y_i))$ by the formula

$$P_n(E)\triangleq\frac{\int_E[r^n(x,\varphi^1(x))-r^n(x,\varphi^2(x))]\,\mu(dx)}{\int_{Y_i}[r^n(x,\varphi^1(x))-r^n(x,\varphi^2(x))]\,\mu(dx)},\quad n=1,2,\dots,N. \tag{3}$$

By Liapunov's theorem, there exists a set $\Gamma_i\in\mathcal{B}(Y_i)$ such that

$$P_n(\Gamma_i)=\int_{Y_i}P_n(dx)\,\hat\pi(x),\quad n=1,2,\dots,N. \tag{4}$$

We define

$$\varphi(x)=\begin{cases}\varphi^1(x), & \text{if } x\in\Gamma_i,\\ \varphi^2(x), & \text{if } x\in Y_i\setminus\Gamma_i.\end{cases}$$

Then

$$P_n(\Gamma_i)=\frac{\int_{Y_i}r^n(x,\varphi(x))\,\mu(dx)-\int_{Y_i}r^n(x,\varphi^2(x))\,\mu(dx)}{\int_{Y_i}[r^n(x,\varphi^1(x))-r^n(x,\varphi^2(x))]\,\mu(dx)},\quad n=1,\dots,N. \tag{5}$$

Since the measures P_n were defined in (3) as differences of two measures, absolutely continuous with respect to the measure μ, Theorem 16.10 in Billingsley [7] implies

$$\int_{Y_i}P_n(dx)\,\hat\pi(x)=\frac{\int_{Y_i}r^n(x,\varphi^1(x))\hat\pi(x)\,\mu(dx)-\int_{Y_i}r^n(x,\varphi^2(x))\,\mu(dx)}{\int_{Y_i}[r^n(x,\varphi^1(x))-r^n(x,\varphi^2(x))]\,\mu(dx)}$$
$$+\frac{\int_{Y_i}r^n(x,\varphi^2(x))[1-\hat\pi(x)]\,\mu(dx)}{\int_{Y_i}[r^n(x,\varphi^1(x))-r^n(x,\varphi^2(x))]\,\mu(dx)},$$
$$n=1,2,\dots,N. \tag{6}$$

From the definition of $\hat\pi$ we have

$$r^n\big(x,\varphi^1(x)\big)\hat\pi(x)+r^n\big(x,\varphi^2(x)\big)\big(1-\hat\pi(x)\big)=\int_{\tilde A(x)}r^n(x,a)\,\pi(da|x).$$

This formula and (4)–(6) imply

$$\int_{Y_i}r^n\big(x,\varphi(x)\big)\,\mu(dx)=\int_{Y_i}\int_{\tilde A(x)}r^n(x,a)\,\pi(da|x)\,\mu(dx),\quad n=1,2,\dots,N. \tag{7}$$

Now we consider the last possible situation when $\mu(Y_i) > 0$ and $\sim_n$ is $=$ for at least one $n = 1, \ldots, N$ in (2). Then for any measurable function φ with $\varphi(x) \in \tilde{A}(x)$ for all $x \in Y_i$, equality (7) holds for all n such that $\sim_n$ is the equality. If $\sim_n$ are equalities for all $n = 1, 2, \ldots, N$, then we may select Γ_i as an arbitrary measurable subset of Y_i. Otherwise, we remove coordinates that correspond to equalities in the definition of Y_i and apply the previous construction of Γ_i to a problem with a smaller number of coordinates. For all of these coordinates, the corresponding relationship in (2) is a strong inequality.

Since formula (7) holds for any Y_i, such that $\mu(Y_i) > 0$ in the finite partition $\{Y_i\}$,

$$\mathbf{R}(P^{\varphi}) = \int_X r\big(x, \varphi(x)\big)\,\mu(dx) = \int_X \int_{\tilde{A}(x)} r(x,a)\,\pi(da|x)\,\mu(dx) = \mathbf{R}(P^{\pi}),$$

as we wished to prove. □

Before we establish the validity of Theorem 2.1 for one-step model, let us prove its simplified version.

Lemma 3.2. *In the one-step model with the nonatomic measure μ*

$$\mathcal{V} \cap \mathbb{R}^N = \mathcal{V}^M \cap \mathbb{R}^N.$$

Proof. Since $\mathcal{V} \supseteq \mathcal{V}^M$, it is sufficient to prove that $\mathcal{V} \cap \mathbb{R}^N \subseteq \mathcal{V}^M \cap \mathbb{R}^N$. According to Feinberg [9, Theorem 5.2], for any fixed $\pi \in \Delta$, there exists a probability measure ν on the set $\mathcal{D}^M$ such that for any $E \in \mathcal{B}(E)$

$$P^{\pi}(E) = \int_{\mathcal{D}^M} Q(E)\,\nu(dQ).$$

Therefore

$$\mathbf{R}(P^{\pi}) = \int_{\mathcal{D}^M} \mathbf{R}(Q)\,\nu(dQ). \tag{8}$$

Suppose that $\mathcal{V}^M \cap \mathbb{R}^N \neq \emptyset$ and $\pi \in \Delta$ is a policy such that $\mathbf{R}(P^{\pi}) \in \mathcal{V} \cap \mathbb{R}^N$. The corresponding measure ν is concentrated on the set of strategic measures Q for which $\mathbf{R}(Q) \in \mathcal{V}^M \cap \mathbb{R}^N$. Therefore, $\tilde{\nu}(\mathcal{V}^M \cap \mathbb{R}^N) = 1$, where $\tilde{\nu}$ is the image of the measure ν under the mapping $\mathbf{R}(\cdot) : \mathcal{D} \to \mathbb{R}^N$. According to Meyer [16, Chapter 2, Theorem 12]

$$\int_{\mathcal{D}^M} \mathbf{R}(Q)\,\nu(dQ) = \int_{\mathcal{V}^M \cap \mathbb{R}^N} r\,\tilde{\nu}(dr).$$

Therefore

$$\mathbf{R}(P^\pi) = \int\limits_{\mathcal{V}^M \cap \mathbb{R}^N} r\,\tilde{\nu}(dr)$$

is the finite expectation of a random variable in $\mathbb{R}^N$ with respect to a probability concentrated on the convex set $\mathcal{V}^M \cap \mathbb{R}^N$; see Lemma 3.1. Thus $\mathbf{R}(P^\pi) \in \mathcal{V}^M \cap \mathbb{R}^N$ and $\mathcal{V} \cap \mathbb{R}^N \subseteq \mathcal{V}^M \cap \mathbb{R}^N$.

Suppose now that $\mathcal{V}^M \cap \mathbb{R}^N = \emptyset$. Then, according to (8), there are no policies for which all the functionals $\mathbf{R}_1(P^\pi), \mathbf{R}_2(P^\pi), \ldots, \mathbf{R}_N(P^\pi)$ are finite. So $\mathcal{V} \cap \mathbb{R}^N = \emptyset$. □

Lemma 3.3. *Let μ be a finite nonatomic measure on X and let $R : X \to [0, \infty]$ be a measurable function on X such that $\int_X R(x)\,\mu(dx) = \infty$. Then there exists a sequence $\{Y_1, Y_2, \ldots\}$ of disjoint measurable subsets of X such that $\int_{Y_i} R(x)\,\mu(dx) > 1$ for all $i = 1, 2, \ldots$.*

Proof. It is sufficient to prove that X can be partitioned into two measurable subsets X_1 and X_2 such that $\int_{X_1} R(x)\,\mu(dx) > 1$ and $\int_{X_2} R(x)\,\mu(dx) = \infty$. Let

$$Z_1 = \bigl\{x \in X\colon\ R(x) < \infty\bigr\}, \qquad Z_2 = \bigl\{x \in X\colon\ R(x) = \infty\bigr\}.$$

First, we consider the case $\mu(Z_2) > 0$. In this case, Lemma 2 in Feinberg and Piunovskiy [10] implies that Z_2 can be partitioned into sets V_1 and V_2 such that $\mu(V_i) > 0$, $i = 1, 2$. We set $X_1 = Z_1 \cup V_1$ and $X_2 = V_2$. Then $\int_{X_i} R(x)\,\mu(dx) = \infty$, $i = 1, 2$.

Second, we consider the case $\mu(Z_2) = 0$. We define $X^K = \{x \in X\colon\ R(x) \leqslant K\}$. Then

$$\int\limits_{X^K} R(X)\,\mu(dx) \nearrow \int\limits_{X} R(X)\,\mu(dx) \quad \text{as } K \to \infty;$$

see, e.g., Neveu [17, Proposition II.3.3]. We select K such that the integral of R over the set X^K is greater than 1. We set $X_1 = X^K$ and $X_2 = X \setminus X^K$.

Since the integral of R over X_1 is finite, the integral over X_2 is infinite. □

Proof of Theorem 2.1 (for a one-step model). First, we observe that Lemma 3.2 is valid also for subprobability measures μ. Second, in order to prove Theorem 2.1 for a one-step MDP, it is sufficient to show that for any policy π there exists a nonrandomized policy φ such that

$$\mathbf{R}(P^\varphi) = \mathbf{R}(P^\pi). \tag{9}$$

Let K^π be the number of coordinates of the vector $\mathbf{R}(P^\pi)$ with infinite values. We prove (9) by induction.

For $K^\pi = 0$, formula (9) follows from Lemma 3.2. Let (9) be valid for $K^\pi = K \geqslant 0$. We prove this formula for $K^\pi = K + 1$.

Without loss of generality we assume that the Nth coordinate of the vector $\mathbf{R}(P^\pi)$ is infinite:

$$R^N(P^\pi) = \int_X \mu(dx) \int_{A(x)} r^N(x,a)\,\pi(da|x) = +\infty. \tag{10}$$

Let $R(x) = \int_{A(x)} r^N(x,a)\,\pi(da|x)$. Then $\int_X R(x)\,\mu(dx) = \infty$. We consider the partition $\{Y_1, Y_2, \ldots\}$ which existence is stated in Lemma 3.3.

We will construct the mapping $\varphi : X \to A$ separately on the sets Y_i, $i = 1, 2, \ldots$. We fix an arbitrary i. There is a positive number M for which

$$\int_{Y_i} \mu(dx) \int_{A(x)} \min\bigl\{r^N(x,a), M\bigr\}\,\pi(da|x) > 1. \tag{11}$$

Since μ is a finite measure, the expression in the left-hand side of (11) is finite. We replace the reward function $r^N(x,a)$ with $\min\{r^N(x,a), M\}$ and apply the induction assumption to the new reward function. We have that there exists a measurable mapping $\varphi : Y_i \to A$ such that $\varphi(x) \in A(x)$ for all $x \in X$ and for all $n = 1, 2, \ldots, N-1$

$$\int_{Y_i} \mu(dx)\, r^n\bigl(x, \varphi(x)\bigr) = \int_{Y_i} \mu(dx) \int_{A(x)} r^n(x,a)\,\pi(da|x) \tag{12}$$

and

$$\int_{Y_i} \mu(dx) \min\bigl\{r^N\bigl(x, \varphi(x)\bigr), M\bigr\}$$

$$= \int_{Y_i} \mu(dx) \int_{A(x)} \min\bigl\{r^N(x,a), M\bigr\}\,\pi(da|x). \tag{13}$$

From (11) and (13) we have that

$$\int_{Y_i} \mu(dx)\, r^N\bigl(x, \varphi(x)\bigr) \geqslant \int_{Y_i} \mu(dx) \min\bigl\{r^N\bigl(x, \varphi(x)\bigr), M\bigr\}$$

$$= \int_{Y_i} \mu(dx) \int_{A(x)} \min\bigl\{r^N(x,a), M\bigr\}\,\pi(da|x) > 1. \tag{14}$$

Since Y_i is an arbitrary element of the partition, function φ is defined on X. Then formula (12) implies

$$R^n(P^\varphi) = R^n(P^\pi), \quad n = 1, 2, \ldots, N-1,$$

and (14) implies that $R^N(P^\varphi) = R^N(P^\pi) = \infty$. □

4. Proof of Theorem 2.1

In this section we extend Theorem 2.1 from one-step to infinite-step MDPs. To do it, we use the construction from Feinberg and Piunovskiy [10] of an equivalent one-step MDP for an infinite-step MDP. We recall that without loss of generality all reward functions are considered to be nonnegative. For $T = 1, 2, \ldots$, we define T-horizon rewards

$$R^n(P^\pi, T) = E^\pi \sum_{t=0}^{T-1} r_t^n(x_t, a_t), \quad n = 1, \ldots, N,$$

and $\mathbf{R}(P^\pi, T) = (R^1(P^\pi, T), R^2(P^\pi, T), \ldots, R^N(P^\pi, T))$.

Lemma 4.1. *For any policy π and for any $T = 1, 2, \ldots$ there exists a randomized Markov policy γ such that*

(i) *γ is nonrandomized at steps $0, 1, \ldots, T-1$;*
(ii) $\mathbf{R}(P^\gamma) = \mathbf{R}(P^\pi)$;
(iii) $\mathbf{R}(P^\gamma, T) = \mathbf{R}(P^\pi, T)$.

Proof. For any policy π, there exists a randomized Markov policy σ such that $P^\sigma(dx_t\, da_t) = P^\pi(dx_t\, da_t)$ for all $t = 0, 1, \ldots$ and therefore $\mathbf{R}(P^\sigma) = \mathbf{R}(P^\pi)$ and $\mathbf{R}(P^\sigma, s) = \mathbf{R}(P^\pi, s)$ for all $s = 1, 2, \ldots$; see Strauch [19, Theorem 4.1]. Therefore, without loss of generality we can assume that π is a Markov policy.

First, we construct a randomized Markov policy which is nonrandomized at epoch 0 and satisfies (ii) and (iii). To do it, we consider a one-step MDP, introduced in Feinberg and Piunovskiy [10, Section 4]. This MDP has the state space X, set of actions $\mathbf{D}$, sets $\mathcal{U}(x)$ of available actions at states $x \in X$, where $\mathbf{D}$ is the set of all strategic measures in the original MDP and $\mathcal{U}(x)$ is the set of all strategic measures in the original MDP such that: (a) the initial distribution is concentrated at x, and (b) the policy is nonrandomized at step 0. By Lemma 8 in Feinberg and Piunovskiy [10], sets $\mathcal{U}(x)$ and the set of all strategic measures $\mathbf{U}$, generated by policies nonrandomized at step 0, are measurable.

We shall prove that the graph of $\mathcal{U}$ is measurable. Let $\mathcal{D}(x)$ be the set of all strategic measures with the initial distribution concentrated at $x \in X$. We consider $\mathcal{D}_0 = \bigcup_{x\in X} \mathcal{D}(x)$ the set of all strategic measures generated by initial distributions concentrated at one point. By Dynkin and Yushkevich [8, Sections 3.6 and 5.5], the sets $\mathcal{D}(x)$ and $\mathcal{D}_0$ are measurable subsets of $\mathbf{D}$.

Let $\mathcal{U}_0 = \bigcup_{x\in X} \mathcal{U}(x)$. Since $\mathcal{U}_0 = \mathcal{D}_0 \cap \mathbf{U}$, the set $\mathcal{U}_0$ is a measurable subset of $\mathbf{D}$. We also observe that $P_x^\pi \to x$ is a measurable projection of $\mathcal{D}_0$ on X see Dynkin and Yushkevich [8, Section 3.6]. Therefore $P_x^\pi \to x$ is a measurable projection of $\mathcal{U}_0$ on X. Thus, the graph of this projection is a measurable subset of $X \times \mathbf{D}$. This graph is the graph of the multifunction $\mathcal{U}(x)$. The measurability of the graph of $\mathcal{U}$ is proved.

For each $x \in X$ and for each $P \in \mathcal{U}(x)$ we consider one-step rewards $\tilde{r}^n(x, P) = R^n(P)$ and $\tilde{r}^{N+n}(x, P) = R^n(P, T)$, $n = 1, \ldots, N$. We have a one-step model with $2N$ criteria $\widetilde{R}^n(\widetilde{P}) = \widetilde{E}\tilde{r}^n(x_0, u_0)$, where $\widetilde{P}$ is a strategic measure in the new one-step model, $\widetilde{E}$ is the expectation in the new model, and u_0 is an action selected in the new model.

We assume that the initial measure μ, which was fixed for the original MDP, is also fixed for the new MDP. Then the new MDP satisfies Condition 2.1.

Lemma 9 in Feinberg and Piunovskiy [10] implies that there is a policy γ in the new MDP such that $\widetilde{R}^n(\widetilde{P}^\gamma) = R^n(P^\pi)$ and $\widetilde{R}^{N+n}(\widetilde{P}^\gamma) = R^n(P^\pi, T)$, $n = 1, \ldots, N$. By Theorem 2.1 applied to the new one-step MDP we have that in this MDP there exists a nonrandomized policy ϕ such that $\widetilde{R}^n(\widetilde{P}^\phi) = \widetilde{R}^n(\widetilde{P}^\gamma)$, $n = 1, \ldots, 2N$. As explained in Feinberg and Piunovskiy [10, p. 62], for the nonrandomized policy ϕ in the new MDP there exists a randomized Markov policy γ^0 such that $R^n(P^{\gamma^0}) = \widetilde{R}^n(\widetilde{P}^\phi)$, $R^n(P^{\gamma^0}, T) = \widetilde{R}^{N+n}(\widetilde{P}^\phi)$, $n = 1, \ldots, N$, and this policy is nonrandomized at step 0. The latter equality follows from the fact that the transformation of ϕ to γ^0 explained in [10] does not depend on particular functions r_t and one can set $r_t^n(x, a) = 0$ when $t \geqslant T$. Therefore, $\mathbf{R}(P^{\gamma^0}) = \mathbf{R}(P^\pi)$ and $\mathbf{R}(P^{\gamma^0}, T) = \mathbf{R}(P^\pi, T)$.

Then we can consider the nonatomic measure $\mu_1(Y) = P_\mu^{\gamma^0}(x_1 \in Y)$. We repeat the previous arguments applied to policy γ^0 on the horizon $1, 2, \ldots$ and construct a Markov policy γ^1 such that it is nonrandomized at the first step and

$$E_{\mu_1}^{\gamma^1} \sum_{t=1}^{T-1} r_t^n(x_t, a_t) = E_{\mu_1}^{\gamma^0} \sum_{t=1}^{T-1} r_t^n(x_t, a_t),$$

$$E_{\mu_1}^{\gamma^1} \sum_{t=1}^{\infty} r_t^n(x_t, a_t) = E_{\mu_1}^{\gamma^0} \sum_{t=1}^{\infty} r_t^n(x_t, a_t)$$

for all $n = 1, \ldots, N$. We define γ^1 at step 0 being equal to γ^0 at that step. Then

$$\mathbf{R}\big(P^{\gamma^1}\big) = \mathbf{R}\big(P^{\gamma^0}\big) = \mathbf{R}(P^\pi),$$

$$\mathbf{R}\big(P^{\gamma^1}, T\big) = \mathbf{R}\big(P^{\gamma^0}, T\big) = \mathbf{R}(P^\pi, T),$$

and the randomized Markov policy γ^1 is nonrandomized at steps 0 and 1.

By repeating this construction $T - 2$ times more, we obtain the policy $\gamma = \gamma^{T-1}$ satisfying conditions (i)–(iii) of the lemma. □

Proof of Theorem 2.1. Fix an arbitrary policy π. To prove the theorem, it is sufficient to show that $\mathbf{R}(P^\sigma) = \mathbf{R}(P^\pi)$ for some nonrandomized Markov policy σ.

Consider a sequence $\epsilon_k \searrow 0$. We define $T_1 > 0$ such that for all $n = 1, \ldots, N$

$$R^n(P^\pi, T_1) \geqslant \begin{cases} R^n(P^\pi) - \epsilon_1, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_1}, & \text{otherwise.} \end{cases}$$

Let γ be a policy which existence was stated in Lemma 4.1 for $T = T_1$. We set $\gamma^1 = \gamma$.

Suppose that for some $k = 1, 2, \ldots$ and for some $T_k \geqslant k$, we have a randomized Markov policy γ^k such that

(a) γ^k is nonrandomized at steps $0, \ldots, T_k - 1$;
(b) for all $n = 1, \ldots, N$

$$R^n\left(P^{\gamma^k}, T_k\right) \geqslant \begin{cases} R^n(P^\pi) - \epsilon_k, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_k}, & \text{otherwise;} \end{cases}$$

(c) $\mathbf{R}(P^{\gamma^k}) = \mathbf{R}(P^\pi)$.

We select $T_{k+1} > T_k$ such that for all $n = 1, \ldots, N$

$$R^n\left(P^{\gamma^k}, T_{k+1}\right) \geqslant \begin{cases} R^n(P^\pi) - \epsilon_{k+1}, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_{k+1}}, & \text{otherwise.} \end{cases}$$

By applying Lemma 4.1 to the MDP with the horizon $T_k, T_k + 1, \ldots$ and with the initial distribution $\tilde{\mu}(Y) = P_\mu^{\gamma^k}(x_{T_k} \in Y)$, we have that there exists a randomized Markov policy γ^{k+1} that satisfies (a)–(c) with k increased by 1. At steps $0, 1, \ldots, T_k - 1$ this policy is defined being equal to γ^k and on steps $T_k, T_k + 1, \ldots$ this policy is constructed by using Lemma 4.1.

We define a nonrandomized Markov policy γ which coincides with γ^k at steps $0, 1, \ldots, T_k - 1$ for all $k = 1, 2, \ldots$. Since $T_k < T_{k+1}$ and $\gamma_t^k = \gamma_t^{k+1}$ for $t < T_k$, $k = 1, 2, \ldots$, this definition is correct. Inequality (b) implies that

$$R^n(P^\gamma, T_k) \geqslant \begin{cases} R^n(P^\pi) - \epsilon_k, & \text{if } R^n(P^\pi) < \infty, \\ \frac{1}{\epsilon_k}, & \text{otherwise,} \end{cases} \tag{15}$$

and equality (c) implies that

$$R^n(P^\gamma, T_k) \leqslant R^n(P^\pi) \tag{16}$$

for all $n = 1, \ldots, N$. Since $\epsilon_k \searrow 0$ and all one-step rewards are nonnegative, (15) and (16) imply $\mathbf{R}(P^\gamma) = \lim_{k\to\infty} \mathbf{R}(P^\gamma, T_k) = \mathbf{R}(P^\pi)$. □

5. Optimization problem

A natural way to study a multicriterion problem is to replace it with a constrained one. Thus we fix finite constants $d_2, \ldots, d_N$ and consider the following optimization problem:

$$\underset{\pi}{\text{maximize}}\ R^1(P^\pi) \tag{17}$$

subject to

$$R^n(P^\pi) \geqslant d_n, \quad n = 2, \ldots, N. \tag{18}$$

Recall that a multifunction $x \to A(x)$ is called upper semicontinuous if, for any open $\Gamma \subseteq A$, the set $\{x\colon\ A(x) \subseteq \Gamma\}$ is open. We consider the following condition.

Condition 5.2. (a) *Either*

(i) *for every $t = 0, 1, \ldots$ the mapping $x \to A_t(x)$ is upper semicontinuous, and $A_t(x)$ is compact for each fixed $x \in X$, or*

(ii) *there exist finite bounded numbers $\{\lambda_{n,t} \mid n = 1, 2, \ldots, N,\ t = 0, 1, \ldots\}$, such that for each $t = 0, 1, \ldots$ the function*

$$r_t^{\lambda}(x, a) = \sum_{n=1}^{N} \lambda_{n,t} r_t^n(x, a)$$

is nonpositive and sup-compact; that is, for any finite number C the level set

$$L_t(C) = \left\{(x, a) \in \mathrm{Gr}(A_t) \mid r_t^{\lambda}(x, a) \geqslant C\right\}$$

is compact.

(b) *The transition probabilities $p_t(dy|x, a)$ are weakly continuous in $(x, a) \in \mathrm{Gr}(A_t)$, $t = 0, 1, \ldots$.*

(c) *The function $r_t(\cdot)\colon X \times A \to \mathbb{R}^N$ is (component-wise) nonpositive and upper semicontinuous, $t = 0, 1, \ldots$.*

Note that under Condition 5.2(a) (i), $\mathrm{Gr}(A_t)$ are Borel and the multifunctions $x \to A_t(x)$ can be uniformized. The first statement follows from Kechris [15, Exersise 25.14] and the uniformization follows from Arsenin–Kunugui theorem; see Kechris [15, Theorem 18.18]. The same is true if Condition 5.2(a) (ii) holds and $r_t^{\lambda}(x, a) > -\infty$ for all $x \in X$ and for all $a \in A_t(x)$. Indeed, in this case, $\mathrm{Gr}(A_t) = \bigcup_{\{l=1,2,\ldots\}} L_t(-l)$ and the uniformization follows from the Arsenin–Kunugui theorem. In particular, $r_t^{\lambda}(x, a) > -\infty$ for all $x \in X$ and for all $a \in A_t(x)$ when $|r_t^n(x, a)| < \infty$ for all $n = 1, \ldots, N$, for all $x \in X$, and for all $a \in A_t(x)$.

The function r_t^{λ} is upper semicontinuous because it is sup-compact. We do not assume that $\lambda_{n,t} \geqslant 0$. If for some $t = 0, 1, \ldots$ we have that $\lambda_{n,t} \geqslant 0$, $n = 1, 2, \ldots, N$, then Condition 5.2(c) implies that the function r_t^{λ} is nonpositive.

A policy π is called feasible if it satisfies (18). If a feasible policy exists, the problem is called feasible too. A feasible policy is called optimal if the maximum in (17) is achieved at this policy.

Theorem 5.1. *If Condition* 5.2 *holds and problem* (17), (18) *is feasible then there exists an optimal policy.*

Proof. We define for $n = 1, 2, \ldots, N$

$$\bar{r}^n(h_\infty) = \bar{r}^n(x_0, a_0, x_1, a_1, \ldots) = \sum_{t=0}^{\infty} r_t^n(x_t, a_t).$$

Condition 5.2(c) implies that the functions $\bar{r}^n(\cdot)$ are nonpositive and upper semicontinuous. Hence, by [18, Lemma 17], the mapping

$$\mathbf{R}(P) = \int_{H_\infty} \bar{r}(h_\infty)\, P(dh_\infty)$$

from $\mathcal{D}$ into $\mathbb{R}^N$ is bounded above and upper semicontinuous.

In case (i), $\mathcal{D}$ is compact [2, Theorem 2.1 and Proposition 3.1], and it remains only to notice that the nonempty set

$$\left\{P \in \mathcal{D}\colon\ R^n(P) \geqslant d_n,\ n = 2, \ldots, N\right\}$$

is compact.

Let us consider case (ii). We observe that without loss of generality the numbers $\lambda_{n,t}$ can be selected in a way that $|\lambda_{n,t}| \leqslant 1$ for all $n = 1, \ldots, N$, $i = 0, 1, \ldots$. Indeed, if any set of $\lambda_{n,t}$ satisfies condition (ii), we can replace all $\lambda_{n,t}$ with $\lambda'_{n,t} = \lambda_{n,t} / \max\{1, \lambda_t\}$, where $\lambda_t = \max\{|\lambda_{n,t}| \mid n = 1, \ldots, N\}$.

Let

$$S = \sup_{P \in \mathcal{D}} \left\{R^1(P)\colon\ R^n(P) \geqslant d_n,\ n = 2, \ldots, N\right\}.$$

If $S = -\infty$ then the assertion of the theorem is trivial. So, let $S > -\infty$. We consider a finite number $d_1 < S$. Clearly, if

$$R^n(P) \geqslant d_n, \quad n = 1, 2, \ldots, N,$$

then

$$R^\lambda(P) \geqslant \sum_{n=1}^{N} d_n,$$

where

$$R^\lambda(P) = \int_{H_\infty} \bar{r}^\lambda(h_\infty)\, P(dh_\infty) \quad \text{and} \quad \bar{r}^\lambda(h_\infty) = \sum_{t=0}^{\infty} r_t^\lambda(x_t, a_t)$$

are nonpositive upper semicontinuous functions. Therefore, problem (17) is equivalent to the following one:

$$R^1(P) \to \sup_{P \in \widetilde{\mathcal{D}}}, \qquad R^n(P) \geqslant d_n, \quad n = 2, \ldots, N, \tag{19}$$

where

$$\widetilde{\mathcal{D}} = \left\{ P \in \mathcal{D} \mid R^\lambda(P) \geqslant \sum_{n=1}^{N} d_n \right\} \neq \emptyset. \tag{20}$$

Now we can prove that problem (19) has a solution. Since all the functionals $R^1(\cdot), \dots, R^N(\cdot)$ are bounded above and upper semicontinuous, it is sufficient to establish that $\widetilde{\mathcal{D}}$ is compact. First of all, note that, for any finite number C, the set

$$E(C) = \left\{h_\infty\colon\ r_t^\lambda(x_t, a_t) \geqslant C \text{ for all } t = 0, 1, \dots\right\}$$

is compact according to Tychonoff theorem [18, p. 292]. Hence, its closed subset

$$\left\{h_\infty\colon\ \bar{r}^\lambda(h_\infty) \geqslant C\right\} = E(C) \cap \left\{h_\infty\colon\ \bar{r}^\lambda(h_\infty) \geqslant C\right\}$$

is also compact. Therefore, the function $-\bar{r}^\lambda(\cdot) \geqslant 0$ is strictly unbounded [14, p. 215]; we recall that the definition of a strictly unbounded function [14, p. 188] covers continuous functions on compact sets because of the agreement that $\inf\{\emptyset\} = \infty$. In view of (20) and [14, Theorem 12.2.15], the set $\widetilde{\mathcal{D}}$ is tight; see also [3, §2]. Therefore, by Prokhorov's theorem ([6] or [14]), this set is relatively compact. It remains to notice that $\widetilde{\mathcal{D}}$ is closed in the space $\mathcal{P}(H_\infty)$ equipped with the weak topology because $\mathcal{D}$ is closed under Condition 5.2(b) (see, e.g., [18, Theorem 9]) and the functional $R^\lambda(\cdot)$ is upper semicontinuous. □

Condition 5.3. *Conditions* 5.2(a) *and* (b) *and the following condition are satisfied*:

(c′) *the function* $r_t(\cdot)\colon X \times A \to \mathbb{R}^N$ *is upper semicontinuous and there exists a sequence* $c_t \geqslant 0$ *such that* $c \triangleq \sum_{t=0}^\infty c_t < \infty$ *and* $r_t^n(x, a) \leqslant c_t$ *for all* $n = 1, 2, \dots, N$ *and for all* $t = 0, 1, \dots, \forall (x, a) \in \mathrm{Gr}(A_t)$.

Corollary 5.1. *If Condition* 5.3 *holds and problem* (17), (18) *is feasible then there exists an optimal policy.*

Proof. We substitute the reward functions $r_t^n(\cdot)$ with $r_t^n(\cdot) - c_t$. The new reward functions satisfy Condition 5.2(c). Let $\overline{R}^n$ be the new objective functions. Then $\overline{R}^n(P) = R^n(P) - c$ for all $n = 1, 2, \dots, N$ and for all policies π. Therefore, problem (17) is equivalent to the similar problem with the new reward functions and with constraints

$$\overline{R}^n(P^\pi) \geqslant d_n - c, \quad n = 2, \dots, N.$$

The latter problem satisfies Condition 5.2 and thus has a solution according to Theorem 5.1. □

Discounted rewards are important applications of Corollary 5.1. For standard discounting, $r_t^n(x, a) = \beta^t r^n(x, a)$, where $\beta \in [0, 1[$ is a constant, and the functions $r^n(x, a)$ are bounded above and upper semicontinuous. In this case, Corol-

lary 5.1 implies Theorem 3.2 in [13]. Corollary 5.1 can be also applied to weighted discounted criteria; see [11]. In this case,

$$r_t^n(x,a) = \sum_{i=1}^{m} \beta_i^t \bar{r}_i^n(x,a),$$

where $m = 2, 3, \ldots$ is some integer, $\beta_i \in [0, 1[$, and functions $\bar{r}_i^n(\cdot)$ are upper semicontinuous and bounded above.

Condition 5.3(c′) also holds when $r_t^n(x,a) = \beta^t \bar{r}_t^n(x,a)$ with the functions $\bar{r}_t^n(\cdot)$ being uniformly bounded above and upper semicontinuous.

Corollary 5.2. *If problem* (17), (18) *is feasible and Conditions* 2.1 *and* 5.2 (*or* 5.3) *are satisfied then there exists an optimal nonrandomized Markov policy.*

Proof. The proof follows directly from Theorems 2.1, 5.1 and Corollary 5.1. □

6. Applications

Example 6.1. Let us consider an inventory system with finite or infinite capacity M. The demand at epoch $t = 0, 1, \ldots$ is ξ_t (nonnegative mutually independent random variables). We assume that the distribution P_t of ξ_t has no atoms and $P_t\{\xi_t < \infty\} = 1$. Orders are placed after the demand is known and it is possible to order up to the full capacity of the system. Let the initial inventory be y. Then the initial state of the system is $x_0 = y - \xi_0$. The dynamics of the system is defined by the equation $x_{t+1} = x_t + a_t - \xi_{t+1}$, $X = (-\infty, M]$, $A_t(x) = [0, M - x]$. (If $M = \infty$ then $X = \mathbb{R}^1$, $A_t(x) = [0, \infty)$.)

Suppose that N-dimensional vectors $r_t(x,a)$ of measurable additive rewards (or losses) with values in $[-\infty, \infty]$ at steps $t = 0, 1, \ldots$ are given. Theorem 2.1 implies that (nonrandomized) Markov policies for this multicriterion problem are as good as general policies.

Let $h_t(x)$ be the holding cost of the amount x during one period of time $[t, t+1)$, and $K_t(a)$ be the ordering cost of a units at epoch t. We assume that for each $t = 0, 1, \ldots$ the functions $h_t(\cdot)$ and $K_t(\cdot)$ are lower semicontinuous, bounded below, and $h_t(x) \to \infty$ as $|x| \to \infty$, $K_t(a) \to \infty$ as $a \to \infty$. We remark that our assumptions cover the following particular functions

$$h_t(x) = \begin{cases} h_1^t x, & \text{if } x \geqslant 0, \\ -h_2^t x, & \text{otherwise,} \end{cases}$$

and

$$K_t(a) = \begin{cases} k_0^t + k_1^t a, & \text{if } a > 0, \\ 0, & \text{if } a = 0, \end{cases} \tag{21}$$

where $h_0^t, h_1^t > 0$ and $k_0^t, k_1^t \geqslant 0$ are some coefficients.

One can consider different reward functions associated with this inventory system. For instance, we may set $r_t^1(x,a) = -\beta_1^t h_t(x)$ and $r_t^2(x,a) = -\beta_2^t K_t(a)$, where $\beta_{1,2} \in (0,1)$ are given discount factors. Then $R^1(\cdot)$ is the criterion characterizing holding and backordering costs, and $R^2(\cdot)$ characterizes operational costs. If we fix appropriate constant d_2 then Conditions 5.2 (version (ii) of item (a) with $\lambda_{n,t} \equiv 1$) are satisfied for the constrained problem (17) with $N = 2$. Thus, there exists an optimal nonrandomized Markov policy for this problem. (See Corollary 5.2.)

For various applications, it is possible to consider different versions of this problem. For example, for discounted problems, instead of ordering costs $K(a)$ one can consider two costs $K'(a)$ and $K''(a)$ where, similarly to (21), the first function is the cost to place an order and the second function is the amount paid for the inventory a. Lower semicontinuity of nonnegative functions h, $K'(a)$ and $K''(a)$, which takes place in applications, implies in view of Corollary 5.2, the existence of optimal nonrandomized Markov policies for various constrained problems of this type.

Example 6.2. An investor has an option to sell a portfolio at epoch $t = 1, 2, \ldots$. The value of the portfolio at epoch $t = 0, 1, 2, \ldots$ is $z_t \in \mathbb{R}^1$. Suppose that $z_0 \geqslant 0$ is given and the value of z_{t+1} is defined by transition probabilities $q_t(dz_{t+1}|z_t)$, $t = 0, 1, \ldots$. We assume that $q_t(\cdot|z_t)$ are nonatomic and weakly continuous.

At each epoch $t = 1, 2, \ldots$, the investor has two options: to sell the whole portfolio or to keep it. We construct a Markov decision process for this problem. Let $X = \{0,1\} \times [0,\infty)$ and $A = \{0,1\}$. Action 0 (1) means to hold (to sell) the portfolio. The state of the system is $x_t = (0, z_{t+1})$ ($x_t = (1, z_{t+1})$) if the portfolio has not been sold (has been sold). In particular, $x_0 = (0, z_1)$ has a nonatomic distribution. For $t = 0, 1, \ldots$ we set $A_t((0,z)) = \{0,1\}$, $A_t((1,z)) = \{0\}$. If at epoch $t = 0, 1, \ldots$ the system is in state $x_t = (0,z)$ and action $a_t = 0$ is selected then the next state is $x_{t+1} = (0,y)$, where y has the distribution $q_{t+1}(dy|z)$. In all other situations, the system moves from state x_t to the state $(1,y)$, where y has the same distribution $q_{t+1}(dy|z)$.

Suppose that N-dimensional vectors $r_t(x,a)$ of measurable additive rewards (or losses) with values in $[-\infty,\infty]$ at steps $t = 0, 1, \ldots$ are given. Theorem 2.1 implies that (nonrandomized) Markov policies for this multicriterion problem are as good as general policies.

We consider the problem when the investor's goal is to maximize the expected gain under the constraint that with at least probability $P > 0$ this gain is greater (or equal) than the given level C. For $t = 0, 1, \ldots$, we define $r_t^1((0,z),a) = \beta^t \cdot az$, $r_t^2((0,z),a) = a \cdot I\{z \geqslant C/\beta^t\}$; $\beta \in (0,1]$ is the given discount factor. Then the problem can be written in the following form:

$$R^1(P^\pi) \to \sup_\pi, \qquad R^2(P^\pi) \geqslant P. \tag{22}$$

Suppose that the number of steps t is limited by T and $q_t(\cdot|z)$ are concentrated on a finite interval $[0, D]$. We set $r_t^n(x, a) = 0$ when $t \geqslant T$. Then we can set $X = \{0, 1\} \times [0, D]$ and Condition 5.3 holds. Therefore, Corollary 5.2 implies that if this problem is feasible then there exists an optimal nonrandomized Markov policy.

Note added in proof

Shortly before this article was published, the authors learned that a one-step nonatomic problem had been studied earlier; see Balder, J. Multivar. Anal. 16 (1985) 260–264 and references therein. When X is a Borel space, the results of Balder's paper can be compared with our results. Being applied to a one-step model, Theorem 2.1 generalizes Theorem 2.3 and Corollary 2.5 from Balder's paper. In addition, a one-step version of Corollary 5.2 is a stronger result than Theorem 1.1 in Balder's paper. In particular, we do not assume condition (1.2) from Balder's paper.

References

[1] E. Altman, Constrained Markov Decision Processes, Chapman & Hall/CRC, Boca Raton, 1999.
[2] E.J. Balder, On compactness of the space of policies in stochastic dynamic programming, Stochastic Process. Appl. 32 (1989) 141–150.
[3] E.J. Balder, Lectures on Young measures, Cahiers de Mathématiques de la Decision, CEREMADE, Université Paris IX, Dauphine, Paris, 1995.
[4] J.R. Barra, Mathematical Basis of Statistics, Academic Press, New York, 1981.
[5] D.P. Bertsekas, S.E. Shreve, Stochastic Optimal Control, Academic Press, New York, 1978.
[6] P. Billingsley, Convergence of Probability Measures, Wiley, New York, 1968.
[7] P. Billingsley, Probability and Measure, Wiley, New York, 1986.
[8] E.B. Dynkin, A.A. Yushkevich, Controlled Markov Processes and Their Applications, Springer-Verlag, New York, 1979.
[9] E.A. Feinberg, On measurability and representation of strategic measures in Markov decision processes, in: T. Ferguson (Ed.), Statistics, Probability and Game Theory, Papers in Honor of David Blackwell, in: IMS Lecture Notes Monograph Ser., Vol. 30, 1996, pp. 29–43.
[10] E.A. Feinberg, A.B. Piunovskiy, Multiple objective nonatomic Markov decision processes with total reward criteria, J. Math. Anal. Appl. 247 (2000) 45–66.
[11] E.A. Feinberg, A. Shwartz, Constrained Markov decision models with weighted discounted rewards, Math. Oper. Res. 20 (1995) 302–320.
[12] E.B. Frid, On optimal strategies in controlled problems with constraints, SIAM Theory Probab. Appl. 17 (1972) 188–192.
[13] O. Hernandez-Lerma, J. Gonzalez-Hernandez, Constrained Markov control processes in Borel spaces: the discounted case, Math. Methods Oper. Res. 52 (2000) 271–285.
[14] O. Hernandez-Lerma, J.B. Lasserre, Further Topics on Discrete-Time Markov Control Processes, Springer-Verlag, New York, 1999.
[15] A.S. Kechris, Classical Descriptive Set Theory, Springer-Verlag, New York, 1995.
[16] P.-A. Meyer, Probability and Potentials, Blaisdell, Waltham, MA, 1966.

[17] J. Neveu, Mathematical Foundations of the Calculus of Probability, Holden-Day, San Francisco, 1965.
[18] A.B. Piunovskiy, Optimal Control of Random Sequences in Problems with Constraints, Kluwer, Dordrecht, 1997.
[19] R.E. Strauch, Negative dynamic programming, Ann. Math. Statist. 37 (1966) 871–890.
[20] D.H. Wagner, Survey of measurable selection theorems, SIAM J. Control Optim. 15 (1977) 859–903.